Stratification in Business and Agriculture Surveys with R

Marco BALLIN (ballin@istat.it)
Giulio BARCAROLI (barcarol@istat.it),
Elena CATANESE (catanese@istat.it),
Marcello D'ORAZIO (madorazi@istat.it),
Italian National Institute of Statistics (Istat)

ABSTRACT

Usually sample surveys on enterprises and farms adopt a one stage stratified sampling design. In practice the sampling frame is divided in non-overlapping strata and simple random sampling is carried out independently in each stratum. Stratification allows for reduction of the sampling error and permits to derive accurate estimates. Stratified sampling requires a number of decisions strictly related: (i) how to stratify the population and how many strata to consider; (ii) the size of the whole sample and corresponding partitioning among the strata (so called allocation). This paper will deal mainly with the problem (i) and will show how to tackle it in the R environment using packages already available on the CRAN.

Keywords: *stratified random sampling; multipurpose surveys, optimisation.*
JEL Classification: *C83*

INTRODUCTION

Traditionally, sample surveys on enterprises and farms are based on probabilistic samples and, in most cases, one stage stratified sampling is applied: the sampling frame is divided in non-overlapping subpopulations called *strata* and sampling is performed independently in each subpopulation. Stratification may give more accurate estimates and permits to derive reliable estimates for each subpopulation of interest (cf. Cochran, 1977, pp. 89-90). In business surveys usually the strata are formed by considering geographical information, type of economic activity and some units' measures of size (e.g. number of employees, turnover, etc.). Similarly, in agriculture surveys the strata can be formed by joining geographical information with the type of farming (specialist crops, specialist livestock, mixed) and some measures of farms' size (e.g. size of areas with crops, livestock, etc.). The variables used for stratification purposes must be observed for all the units in the target population and normally can only be chosen among the ones available in the sampling frame; the higher is the association/correlation between the auxiliary

and the target variables, the higher will be the benefits in using them for stratification purposes.

Stratification of a population is relatively simple and straightforward when performed on categorical variables, such as geographical regions, type of activity, etc.; on the contrary if the auxiliary variables are continuous then their categorization is required. Multipurpose surveys, aimed at investigating more phenomena simultaneously, pose a number of problems; as far as stratification is concerned, the main issue is that strata being homogenous with respect to one phenomenon may be heterogeneous with respect to another one.

This paper tackles the problem of stratifying the target population in business and agriculture surveys when the sampling frame provides a set of continuous auxiliary variables available for that purpose. Univariate and multivariate methods will be presented together with the corresponding R packages implementing them. Main features and notation of stratified random sampling design are provided in Section 2. Section 3 deals with the problem of stratification in presence of a single continuous variable, providing also hints for performing the task in the R environment. Section 4 tackles the problem of multivariate stratification, by presenting the approach developed at Istat and the associated R package. Section 5 provides an example of application of the various strategies when designing the Farm Structure Survey, aimed at investigating the main structural characteristics of the agriculture holdings.

STRATIFIED SAMPLING

In stratified random sampling, the sampling frame is divided in non-overlapping subpopulations or *strata* and sampling is performed independently in each stratum. In formal terms, if U is the finite population under investigation, consisting of N units, for sampling purposes it: (1) is divided in H non-overlapping strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$), whereas N_h denotes the number of units in the stratum h and, consequently, $N = \sum_{h=1}^{H} N_h$; then, (2) a sample s_h of n_h ($n_h \leq N_h$) units is selected in the hth stratum ($h = 1,2,\ldots,H$) and the overall sample size is obtained as $n = \sum_{h=1}^{H} n_h$.

The decisions to take concern mainly: (i) how to stratify the population and how many strata to create; (ii) the probabilistic criterion to select the sample in each stratum; and (iii) the size of the whole sample and corresponding partitioning among the strata (so called *allocation*). The topic (i) will be tackled in Section 3. As far as point (ii) is concerned, stratification allows for different independent selection schemes in each stratum (simple random sampling, systematic, probability proportional to size, etc.); widespread practice in business and agriculture survey is to apply simple random sampling without replacement (SRSWOR) in all the strata because of its practical and

theoretical advantages. The SRSWOR ensure equal inclusion probabilities, $\pi_{hk} = n_h / N_h$ ($k = 1,2,\ldots,N_h$; $h = 1,2,\ldots,H$) to all the units belonging to the same stratum, while they can change stratum by stratum, unless a constant sampling fraction is considered, i.e. $\pi_{hk} = n_h / N_h = f_0$ ($k = 1,2,\ldots,N_h$; $h = 1,2,\ldots,H$). Such a special case corresponds to a *proportional allocation* of the whole sample among the strata. Alternative allocation criteria may be employed: *equal allocation*, *Neyman allocation*, *power allocation* etc. As far as the overall sample size is concerned, usually it depends mainly on the desired precision for the main survey estimates (expressed in terms or *relative standard error*: desired sampling error divided by the quantity to estimate, denoted usually as Coefficient of Variation, CV). In some surveys the maximum values of CVs in estimating the total amount for a set of variables are explicitly mentioned in European Union (EU) regulations concerning the survey itself.

Estimation of totals in stratified random sampling is straightforward; in particular, if *Y* is the target variable and y_k is its value for the *k*th unit, then an estimate of its total amount $t_y = \sum_{k \in U} y_k$ in *U* is provided by:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k \qquad [1]$$

and the corresponding sampling variance is:

$$V\left(\hat{t}_y\right) = \sum_{h=1}^{H} V\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) \frac{S_{yh}^2}{n_h} \qquad [2]$$

Being S_{yh}^2 the variance of *Y* in the *h*th stratum. As mentioned before, the sampling variance can be expressed in relative terms by computing the *relative standard error* or coefficient of variation $CV_{t_y} = \sqrt{V\left(\hat{t}_y\right)} \Big/ t_y$.

By fixing in advance the desired maximum value of the relative error *d* in estimating the total amount of *Y,* it is possible to determine the required sample size:

$$n_{opt} = \frac{\sum_{h=1}^{H} \frac{N_h^2 S_{yh}^2}{f_h}}{t_y^2 d^2 + \sum_{h=1}^{H} N_h S_{yh}^2} \qquad [3]$$

Where $f_h = n_h / n$ is the proportion of the sample units allocated in the *h*th stratum. In practice, to determine the optimal sample size it is necessary to define the allocation rule. The simplest choice corresponds to *proportional allocation*, $f_h = N_h / N$ ($n_h = n_{opt} N_h / N$); alternatively, it is

possible to resort to *X-proportional allocation* where $f_h = t_{xh} / t_x$ more suited to deal with skewed distributed variables. In *optimal allocation*, often called *Neyman allocation*, the sampling fraction is higher for more heterogeneous strata having $n_h \propto N_h S_h$. *Power allocation* (cf. Särndal et al., 1992, pp. 470-471) sets $n_h \propto t_{yh}^a$ with $0 < a \leq 1$; it is a compromise between Neyman and an allocation ensuring constant precision for each of the strata estimates, avoiding underrepresentation of small strata.

The derivation of the optimal sample size and the corresponding allocation between the strata require information concerning the target *Y* variable, usually not known in advance. For this reason it is substituted by a proxy variable *X*, known for all the units in the population and supposed to be highly correlated with *Y*; in practice, the optimal sample size can be estimated by using the expression [3] whereas all the parameters related to *Y* are substituted by the corresponding ones computed for *X*.

In multipurpose surveys a unique sample should satisfy precision requirements concerning several target variables; in this case the decisions concerning the overall sample size and the corresponding allocation can be approached in a multivariate setting by expressing it as a convex mathematical programming problem. Bethel (1989) and Chromy (1987) provided a solution to this problem by defining an efficient algorithm which always finds the best solution. This algorithm in implemented in "bethel" R package (De Meo, 2012) and in the bethel function in "SamplingStrata" R package (Barcaroli *et al*, 2016).

Obviously, the decision concerning the sample size and its allocation among the strata can be performed only after the stratification of *U*, i.e. given *H* and the corresponding partitioning of *U* in non-overlapping strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$).

STRATIFICATION OF THE TARGET POPULATION

Univariate stratification

An efficient stratification should create strata as homogeneous as possible for the target variable *Y*. Unfortunately the variable *Y* is not known in advance and consequently the stratification is carried out on one or more auxiliary variables strictly related with the *Y*. The stratification does not pose problems if *X* is a categorical variable (e.g. Regions, NACE in case of business surveys or Farm Typology in agriculture, etc.); on the contrary, when *X* is a continuous variable it is necessary to categorize it. In literature different methods are available. A well-known approach is the Dalenius and Hodges (1959) *cumulative* $\sqrt{f}$ *rule*. In particular, once decided the desired number

of strata, *H*, the method identifies the stratum boundaries so that each stratum accounts for $1/H$ of the total integral of $\sqrt{f(x)}$. Opportune approximations are considered to account for the finite population framework.

The cumulative $\sqrt{f}$ rule is not suitable for variables showing a highly skewed distributions; a typical situation in most business and agriculture surveys, where most of the auxiliary and target variables present high positive skewness. A possible strategy in these cases consists in identifying the largest units in terms of *X* values and separating them in a specific stratum, while the remaining units can be stratified by the *cumulative* $\sqrt{f}$ *rule*. The largest units grouped in a unique stratum are included in the sample with certainty (so called *take-all* stratum; Hidiroglou and Lavallée, 2009); in practice, given that these units contribute at large extent to the total amount of the target population, separating them in a stratum that is censused allows to decrease the whole sample size. Hidiroglou (1986) proposed an iterative algorithm to identify the threshold b_c such that all the units with *X* values exceeding it ($x_k > b_c$) are put in the take-all stratum; the procedure requires the specification of the desired CV. Lavallée and Hidiroglou (1988) introduced a unified procedure for both identifying the take-all stratum and stratifying the remaining units. The procedure starts by specifying the desired level of precision (CV) and ends up with a stratification that minimizes the overall sample size. The allocation is based on the power allocation criterion. The Lavallée and Hidiroglou method is based on an iterative procedure which unfortunately may not converge, however convergence problems can partly be solved by using the procedure suggested by Kozak (2004). The geometric stratification method proposed by Gunning and Horgan (2004) is suitable to stratify skewed populations and does not suffer of convergence problems, but it does not separate large units in a take-all stratum and requires the specification of an allocation criterion. As noted by Hidiroglou and Lavallée (2009) it can be a good starting point of the Lavallée-Hidiroglou procedure.

To overcome the problem of working with a variable, *X*, that does not correspond to the target one (*Y*), Rivest (2002) suggested using anticipated moments of *Y* given *X* in the Lavallée and Hidiroglou procedure. Recently, Baillargeon and Rivest (2009) proposed a modification of the procedure by introducing the possibility of separating small units in a stratum that is not sampled (*take-none* stratum), as small units could have a negligible contribution to the total amount of the interest variable, which usually holds true in presence of highly positive skewed distributions. The same authors provided an important contribution to the applicability of the various above mentioned methods by developing the software package "stratification" (Baillargeon and Rivest, 2011, 2014) freely available for the R environment (R Core Team, 2015).

In particular the package "stratification" provides the function `strata.cumrootf` to apply *cumulative* $\sqrt{f}$ *rule*; this function is designed to perform the stratification given the desired *H* strata and permits to derive the sample size given a CV or, vice versa, the CV is computed given a value for the total sample size. Moreover the sample is allocated among the input strata according to the desired allocation criterion. Similarly, `strata.geo` implements Gunning and Horgan geometric stratification. Finally, the function `strata.LH` implements the generalized Lavallée-Hidiroglou (1998) method. All the above mentioned functions allow to model the relationship between the available stratification variable *X* and the unknown target variable *Y*; `strata.LH` handles different models of *Y* vs. *X* (an heteroscedastic linear model and a random replacement model as in Rivest, 2002) and methods; by default the optimization problem is solved by means of the Kozak algorithm, in alternative it is possible to apply the Sethi algorithm. In addition the function permits to identify, if required, the take-none, the take-all strata or both. The following example shows how to call `strata.LH` in order to create $H = 5$ strata (argument Ls) with the last one being the take-all one (`takeall` = 1), when $\mathrm{CV} \leq 0.05$ (argument CV), and allocation following the Neyman criterion (argument `alloc`).

```
strata.LH(x, CV = 0.05, Ls = 5, takeall = 1,
          alloc = list(q1 = 0.5, q2 = 0, q3 = 0.5))
```

Multivariate stratification: the genetic algorithm based method

Given a population frame where a set of auxiliary variables are available, and a set of precision constraints (CVs) associated to its target estimates, the corresponding best stratification can be obtained by adopting a procedure that is based on an exhaustive exploration of the universe of possible solutions. This procedure consists of the following steps:

1. determine the most detailed stratification (atomic strata) by cross-classifying units in the sampling frame using all the values of available auxiliary variables (after categorizing continuous ones);
2. in each atomic stratum of this initial stratification, calculate distributional parameters (mean and variance) of target variables if information related to *Y*s is available for each unit in the frame (or, if not, by using proxy information by other sources);
3. generate all possible partitions from the set of atomic strata;
4. for each generated partition: (a) calculate distributional parameters (mean and variance) of target variables in current strata by aggregating the corresponding information available in atomic

strata; and (b) solve the allocation problem for the current partition and associate the cost of the solution;
5. choose the best stratification as the one given by the partition with the minimal associated cost.

Unfortunately, an exhaustive enumeration in most cases is not feasible, as the number of generated partitions grows exponentially with respect the number of atomic strata.

Package "SamplingStrata" allows to explore the space of stratifications without being obliged to enumerate it exhaustively, by using the genetic algorithm (GA), a search technique that make use of concepts derived by biology, such as population, individual, genome, evolution, fitness, selection, crossover, mutation. (DeJong, 2006).

Under this approach, a given stratification is considered as an *individual* in a population (or *generation* of individuals); an individual is characterized by a *genome* that is optimized in the course of the *evolution*; the genome is represented by a vector whose dimension is given by the number of atomic strata (K) and where to each position in this vector is associated a given atomic stratum; to each element in the vector is assigned randomly an integer value lying in the interval (1,K): atomic strata that share the same integer value, collapse in an aggregate stratum; the *fitness* of each individual is evaluated by using the Bethel algorithm; in the passage from one generation to the next, a percentage of those with higher fitness are directly moved to the next generation (*elitism*), the others are subject to *selection*, i.e. they are randomly selected with probability proportional to their fitness, in order to let them procreate *children*; each child is procreated by applying *crossover* to their parents (a swap of the genes contained in the two genomes), and applying *mutation* to the resulting genome.

At the end of the evolution (the chain of generated populations), the individual with the absolute best fitness will be chosen as the final solution: the genome of this individual represents a stratification in which all or some of the atomic strata have been aggregated, and to which the application of the Bethel algorithm determines a total cost of the sample that is considered the minimum.

The function that performs the optimization of the strata is optimizeStrata:

```
library(SamplingStrata)
data(errors)
data(strata)
# optimisation of sampling strata
solution <- optimizeStrata (
    errors = errors,
    strata = strata,
    cens = NULL,
    strcens = FALSE,
    initialStrata = nrow(strata),
    addStrataFactor = 0.01,
    minnumstr = 2,
    iter = 500,
    pops = 20,
    mut_chance = 0.05,
    elitism_rate = 0.2,
    highvalue = 100000000,
    suggestions = NULL,
    writeFile = FALSE,
    showPlot = TRUE)
sum(ceiling(solution$aggr_strata$SOLUZ))
head(solution$aggr_strata)
```

where the most important parameters are the number of iterations (`iter`), the population size (`pop`), the mutation rate (`mut_chance`), the minimum number of units per stratum (`minnumstr`).

The package "SamplingStrata", together with this optimization function, provides also a number other functions, useful to perform all necessary activities:

- automatic generation of all the information related to atomic strata, starting from the sampling frame (`buildStrataDF`);
- tuning of parameters for the optimization process by varying a number of them in a controlled mode (`tuneParameters`);
- correspondence between atomic strata and optimized strata (`updateStrata`);
- attribution to each unit in the frame of the label indicating the stratum to which the unit belongs (`updateFrame`);
- selection of the sample on the basis of the allocation calculated with respect to optimized strata (`selectSample`).

The method and its implementation in "SamplingStrata" are fully described in Barcaroli (2014).

APPLICATION TO FARM STRUCTURE SURVEY AND COMPARED EVALUATION

The methods proposed in the previous Sections led to define two alternative strategies to tackle the problem of stratifying the population and then allocating the sample:

A. Traditional strategy based on (a.1) stratification of the population based on the cross-classification of units according to few variables, whereas continuous variables are categorized according to the generalized Lavallée-Hidiroglou procedure; and (a.2) calculation of the optimal sample size and corresponding allocation based on the Bethel's method, given the desired CVs.
B. New strategy based on the application of the genetic algorithm, introduced by Ballin and Barcaroli (2013), which performs jointly stratification and allocation.

Both the strategies can easily be implemented in R by using the functions made available by the "stratification" and "SamplingStrata" packages. In this Section the two strategies are compared when applied to design the sample of the Farm Structure Survey (FSS), a very important survey on agriculture holdings carried out every three years (in the Census occasion it is substituted by the Census itself) which should provide estimates of a wide set of characteristics of the farms at national and regional level (NUTS 2). The estimates for the most important crops or livestock characteristics in a given region (identified using criteria established by EU Regulations) should come with a relative standard error not exceeding the 5% (with the exception of small regions). The target population of the FSS excludes the smallest agricultural holdings which together contribute 2 % or less to the total *utilized agricultural area* (UAA) and 2 % or less to the total number of farm *livestock size units* (LSU) obtained as a linear combination (coefficients established by EU Regulation) of the different species of livestock. For sake of simplicity, in this paper it is considered just "Veneto" region, where, according the FSS Regulation, many farm characteristics should be investigated (for details see Table 1). The target population in Veneto consists of 119,384 farms, the sampling frame is built by using the 2010 Agriculture Census results.

Auxiliary and target variables used for stratification and allocation purposes

Table 1

Y variables	CVs	Variables *X* used for stratification	
		Set 1	Set 2
UAA	0.04	UAA	UAA
Cereals	0.05	LSU	Cereals
Oil seed crops	0.05		Industrial crops
Harvested green	0.05		Harvested green
Permanent grassland	0.05		Permanent grassland
Vineyards	0.05		Vineyards
LSU	0.04		LSU
Dairy cows	0.05		Dairy Cows
Other bovines	0.05		Other bovines
Pigs	0.05		Pigs
Poultry	0.05		Poultry

Table 1 reports the variables used for stratification purposes. In particular, the Set 1 is defined according to the principle of parsimony, including only the UAA (observed in 2010 census occasion), and the LSU computed using the 2010 census data are considered. As far as Set 2 is concerned, we consider as stratification variables all the target variables, whose values are the ones observed in the 2010 Census.

First, we apply Strategy A, in particular Kozak implementation of the Lavallée-Hidiroglou method is considered, given its best performances and higher flexibility according to the same authors. Table 2 provides results obtained by considering different initial categorizations for the two auxiliary variables in Set 1. Two different situation are considered, the first without a take-all stratum and the second with a take-all stratum formed by all the strata whose units belong to the last category of one of the stratification variables (i.e. in the first row of Table 2 all the farms in the 5th category of UAA or in the 5th category of LSU are grouped in a take-all stratum).

Optimal sample size with stratification obtained as combination of univariate stratifications (strategy A)

Table 2

Categories for UAA	Categories for LSU	No take-all stratum		With a take-all stratum		
		H	*n*	*H*	Overall *n*	*n* take-all
5	5	25	3358	23	3385	62
6	6	34	3140	34	3217	42
7	7	47	2992	47	3065	36
8	8	61	2880	60	2848	20
9	9	77	2819	76	2862	17
10	10	95	2715	93	2752	15
12	12	136	2638	126	2647	9
15	15	202	2603	199	2631	7
20	20	352	2656	336	2652	6

Table 2 and Figure 1 show that the increase in the number of strata determine a decrease of the overall sample size, but with a very high number of strata the tendency reverses. This is an expected result, in fact creating too many strata can lead to a marked increase of the sample size, also because at least two sample units per stratum are selected. For fixed *H*, the presence of a take-all stratum sometimes determine a slight increase of the overall sample size.

It is worth noting that the computational effort required to perform generalized Lavallée-Hidiroglou stratification, as implemented in the strata.LH function, increases with the number of desired strata. When the desired number of strata is greater than 20 (argument Ls in strata.LH), it is not possible to perform the task with the platform available in our experiments (8GB RAM with Windows 7 OS).

Sample size, take-all units, atomic and final strata using univariate method in Strategy A

Figure 1

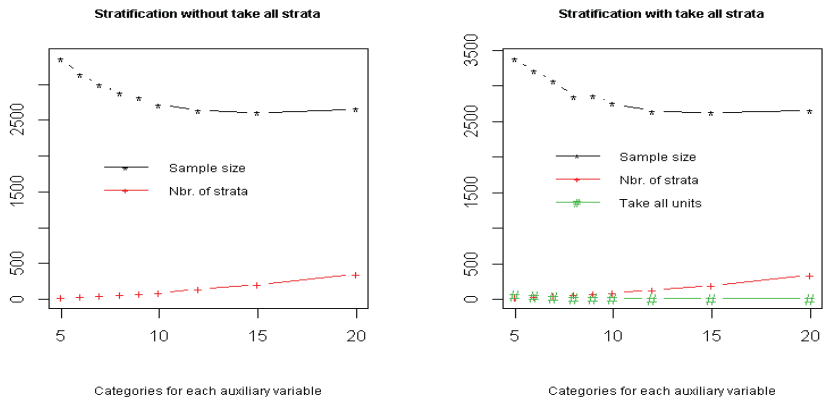


Table 3 and Figure 2 provide results obtained applying strategy B with the two stratification variables in Set 1.

Again, the required sample size decreases with the increase in the number of initial (atomic strata) resulting from the cross-classification of the two categorized auxiliary variables in Set 1 (the categorization is always obtained by applying the *k*-means algorithm). The tendency reverses with too many initial strata. Actually, this result is not to be expected in absolute terms, as it depends on the number of iterations allowed: if we increase the iterations, the sample size should tend at least to replicate the already obtained minimum value. Number of iterations has been limited to 5,000 because GA execution is

quite time consuming. With this limitation, the minimum value of the overall sample size is 2,472, slightly better (-5.0%) than the one obtained by applying the Strategy A (2,603) with the same set of stratification variables.

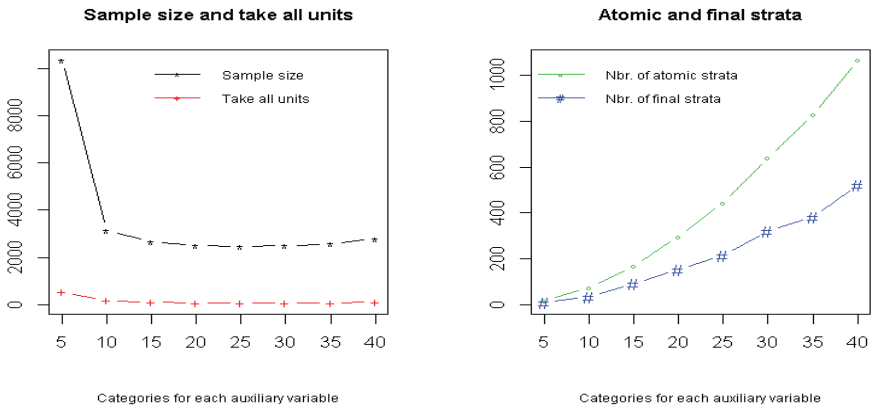
Optimal sample size with stratification obtained by applying GA method (strategy B)

Table 3

Categories for UAA	Categories for LSU	Atomic strata	Optimized strata *H*	Overall *n*	of which *n* take-all
5	5	19	10	10356	537
10	10	74	36	3132	186
15	15	169	91	2680	112
20	20	298	155	2507	89
25	25	443	213	2472	73
30	30	642	321	2495	70
35	35	828	385	2567	76
40	40	1066	522	2795	122

Sample size, take-all units, atomic and final strata using GA method in Strategy A

Figure 2



It is worth noting that the application of the GA method to the categorization of the two auxiliary variables resulting by the univariate method, always produces an improvement of the univariate solution. For instance, consider the best solution obtained by the univariate method, that is the one with a number of classes set to

15 (202 strata and a sample size of 2,603). The GA method receiving in input the 15 categories of UAA and LSU determined by the univariate method, produces a solution characterized by only 106 strata and a sample size of 2,541. The reduction in terms of number of units is not relevant (only -2.5%), but the reduction in terms of number of strata is quite important, as it is almost halved.

Applying strategy A starting from all the variables in the set 2, is unfeasible because the overall number of strata would easily explode. For instance, by categorizing each of the stratification variables in the Set 2 in just 3 categories would determine H = 1,048 non-empty strata and the corresponding optimal sample size would be n = 2,090, a very small value (the smallest if compared to the values presented in tables 2 and 3) which in practice means taking in average two sample units per stratum. Such a situation would be difficult to be managed in practical situations even because nonresponse may turn out with no responding units in a high number of strata requiring an additional effort to compensate for it.

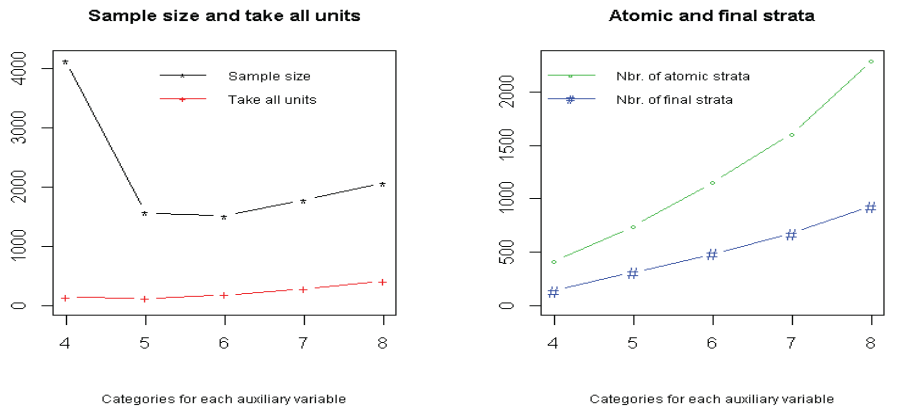
Optimal sample size with stratification obtained by using the Genetic Algorithm (GA) based method (strategy B)

Table 4

Categories for each of the *X* variables	Atomic strata	Final Strata (*H*)	Sample size (*n*)
4	418	140	4132
5	740	312	1571
6	1153	486	1516
7	1606	674	1784
8	2290	931	2065

Sample size, take-all units, atomic and final strata using GA method in Strategy B

Figure 3



Finally, Table 4 and Figure 3 provide results of application of Strategy B when starting with the stratification variables in set 2. When only 4 categories are considered for each of the available *X* variables in the set 2, the procedure ends with 140 non-empty strata and an optimal sample size of 4,132 greater than sample sizes obtained by applying the strategy A with the Set 1. The overall sample size necessary to achieve the target CVs dramatically decreases to 1,571 when considering 5 categories for each of the *X*s. This sample size is smaller than all the results obtained in strategy A with set 1, but the price to pay is that of having a higher number of final strata (312). Results can still be further improved by considering an initial finer stratification (6 categories for each *X*); but if a deeper initial stratification is considered the sample size tends to increase (for the same reason explained before). The results presented in Table 4 are based on a run of 5,000 iterations (argument iter in the function optimizeStrata) carried out for each different starting initial stratification. The processing time required is high: for instance, the elapsed time required to run the optimization step producing a sample size of 1,516 is 33,279.64 seconds on a Windows 7 platform (Intel Pentium 64 bit 2.60 Ghz and 16 GB of RAM).

CONCLUSIONS

As shown in the previous sections, the R environment offers different alternatives to design agriculture or business sample surveys; in particular this work concerns stratified sampling design in presence of continuous auxiliary variables and sample allocation. The paper focuses on the comparison of two different strategies to tackle the problem of categorizing continuous auxiliary variables of the sampling frame. The Lavallée-Hidiroglou procedure allows to choose the allocation criterion, as well as if the last stratum must be take-all, and the take-none stratum. In many multipurpose surveys, such as in agriculture, the stratification based on a unique continuous auxiliary variable may not be adequate. The traditional approach, denoted as strategy A, consists of an exhaustive analysis so to choose a small set of auxiliary variables to be categorized by means of the univariate stratification procedure; then the final strata are derived by cross-classification and the allocation problem is solved by applying the Bethel algorithm to resulting strata. The second approach is based on the exploration of possible stratifications resulting by the cross-classification of all available auxiliary variables, choosing the one that corresponds to minimum cost associated to the sample necessary to satisfy precision constraints (also in this case determined by the Bethel algorithm). In this latter approach the problem of stratification and allocation are solved jointly in a unique step, which makes it innovative. This new approach allows the designer to use a priori all

the available auxiliary information, without caring of possible redundancies to define the initial atomic partition of the population. As the algorithm offers many different parameters, some attempts are required in order to achieve the desired results.

In this paper the two strategies are compared when applied to the design of the Farm Structure Survey. The comparison shows that the innovative strategy should be preferred when the priority is minimizing the overall sample size; the traditional one, which generally performs worse in terms of sample size, may be preferable when the designer wants to keep the total number of strata low and to easily interpret them, or when the designer wants only large units to be in a take-all strata. It is worth noting that the innovative algorithm achieves the best result (1,525) at cost of an average of 5 units per sampled strata, while the traditional approaches achieves its best results in terms of sampled units (2,586) but with an average of roughly 25 sampled units per strata. It must be pointed out that the Genetic Algorithm allows to keep the total number of strata low even if this does not decrease the total computation time nor allow to easily interpret strata. Of course while the traditional approach needs many subjective choices of the designer, the GA automatically explores the space of all the possible partitions of the population thus achieving an optimal solution in terms of sample size that takes into account more complex relationships between auxiliary variables.

A final remark regards a possible joint use of the two methods. This can be seen from two different point of views. First, consider a solution obtained with strategy A: it can be improved by giving it as an input to the GA algorithm: in general, there should be an improvement in terms of sample size, and a greater one in terms of number of strata. Second, if we consider strategy B, as the GA requires a previous step in order to categorize continuous variables, instead of using the *k*-means method it is possible to apply the Lavallée-Hidiroglou method, which in many situations permits to obtain better final results.

References

1. **Baillargeon S** and **Rivest L.-P.,** 2009, *A general Algorithm for Univariate Stratification*. International Statistical Review, 77: 331-344.
2. **Baillargeon S** and **Rivest L.-P.**, 2011, *The Construction of Stratified designs in R with the package stratification*. Survey Methodology, 37: 53-65.
3. **Baillargeon S** and **Rivest L.-P.,** 2014, *Stratification: Univariate Stratification of Survey Populations*. R package version 2.2-5. ***http://CRAN.R-project.org/package=stratification***
4. **Ballin M** and **Barcaroli G.,** 2013, *Joint Determination of optimal Stratification and Sample Allocation Using Genetic Algorithm,* Survey Methodology, 39: 369-393
5. **Barcaroli G.**, 2014. *SamplingStrata: An R Package for the Optimization of Stratified Sampling*. Journal of Statistical Software, 61(4), 1-24. ***http://www.jstatsoft.org/v61/i04/***

6. **Barcaroli G., Pagliuca D., Willighagen E.** and **Zardetto D.**, 2016, *SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys*. R package version 1.1 ***https://CRAN.R-project.org/package=SamplingStrata***
7. **Cochran, W.G.,** 1977, *Sampling Techniques, 3rd Edition*, John Wiley & Sons, New York.
8. **Dalenious T.** and **Hodges J.L.** 1959, *Minimum variance Stratification*. Journal of the American Statistical Association, 54: 88-101.
9. **DeJong K.A.**, 2006, *Evolutionary Computation: a Unified Approach*. MIT Press, Boston, MA
10. **Hidiroglou M.A.**, 1986, *The construction of a self-representing stratum of large units in survey design.* The American Statistician, 40: 27-31.
11. **Hidiroglou M.A.** and **Lavallée P.,** 2009, *Sampling and Estimation in Business Surveys*, in Sample Surveys: Design, Methods and Applications, Vol. 29A, Elsevier
12. **Kozak M.**, 2004 *Optimal Stratification Using Random Search Method in Agricultural Surveys,* Statistics in Transition, 6: 797-806.
13. **Lavallée P.** and **Hidiroglou M.A.**, 1988, *On the Stratification of Skewed Populations.* Survey Methodology, 14: 33-43.
14. **Rivest L.-P.**, 2002, *A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys.* Survey Methodology, 28: 191-198.